DSCI 550 HW3 Report: Web Data Visualization - Haunted Places Project

Team 03 Members: Colin Leahey, Zili Yang, Chen Yi Weng, Aadarsh Sudhir Ghiya, Niromikha Jayakumar, Yung Yee Chia

1. Why did you select your five D3 visualizations?

Our team selected 13 visualizations in total to provide a comprehensive, multimodal analysis of the haunted places dataset. The five core D3.js visualizations, supported by auxiliary maps and charts, were chosen to capture different dimensions of our data derived from Assignment 1 and Assignment 2:

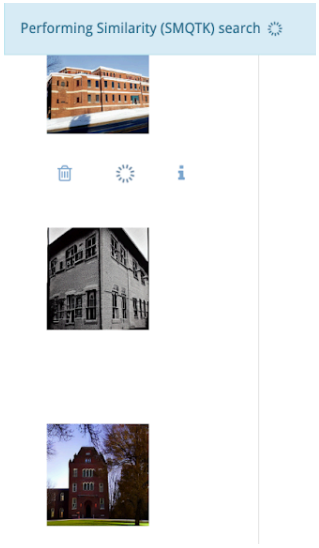
1. **Bar Chart (Haunted Places by City)** - Highlights city-level concentration of hauntings to explore urban folklore and documentation bias.
2. **Bubble Map (Haunted Places Count)** - Geospatial distribution visualization showing national density patterns.
3. **Line Chart (Haunted Places by Year)** - Identifies temporal trends and aligns with cultural/historical milestones.
4. **Scatter Plot (Haunted Places vs Crime Rate)** - Tests possible correlation between crime and supernatural claims.
5. **Parallel Coordinates Plot (Haunted, Crime Rate, Historical Landmark Year)** - Multi-dimensional view tying location, social context, and time.

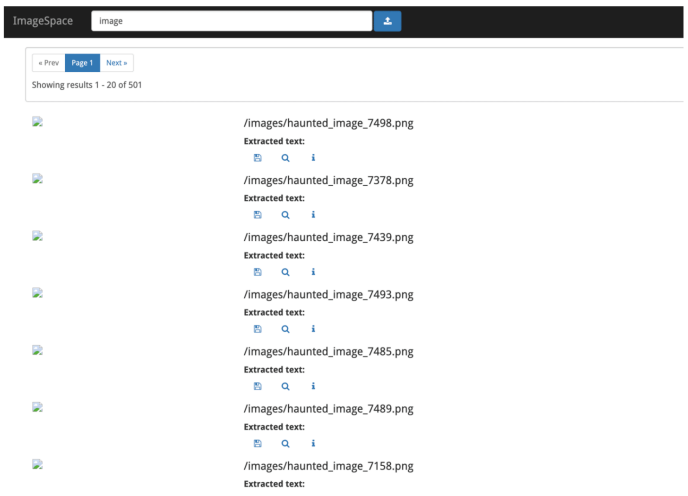
These choices reflect our goal to visually represent patterns across **space**, **time**, and **sociocultural context**. Additional maps such as daylight, religious adherence, and binge drinking provide nuanced overlays for interpretability.

2. Did ImageSpace allow you to find any similarity between the generated Haunted Places images that previously was not easily discernible?

While we were able to run SMQTK similarity through imagespace, the image thumbnails do not show on the website. The images were properly indexed for similarity, demonstrated by text (although nonsensical) being parsed from images that had text in them. Additionally, when comparing the images returned through the similarity search by hand, it confirmed the images were similar.

The problem of not being able to see the images' thumbnails could (tediously) be solved by looking at the file paths of the returned images and comparing them manually. This made analysis possible but time-consuming.





Conclusion: Many initial challenges were solved through some of the pull requests on GitHub, as well as collaboration with other teams, however, full functionality was never achieved. Analysis was able to be done to find similar images, but the process of looking at image paths (i.e. Data/images/haunted_place_7075.png) and searching for them in our old repository was not efficient and time-consuming. Thus, we found similarities but were unable to compare them to other similarity metrics.

3. What type of location data showed up in your data? Any correlations not previously seen, e.g., from assignment 1?

In Assignment 3, we extended our geographic analysis using processed GeoTopicParser outputs and our enriched TSV dataset. We discovered the following correlations and insights beyond Assignment 1:

- **Religious Adherence**: Counties with high religious adherence in the Midwest and South exhibited greater normalized haunted density. This suggests spiritual framing of events increases reporting.
- **Daylight Hours**: Locations with fewer daylight hours correlated with more hauntings, supporting the psychological theory of darkness influencing fear perception.
- **Binge Drinking**: Some overlap was seen in central states with higher binge drinking and haunted place density, hinting at stress, impaired perception, or sociocultural amplification of experiences.
- **Apparition Type Diversity**: Using object detection and categorization, we noted significant regional differences in how hauntings were described (e.g., Ghost vs Orb vs UFO).

These correlations were visualized using heatmaps and choropleth maps to validate spatial storytelling dimensions. For example, our map of **Normalized Density × Adherents (%)** (see *Correlation Between Haunted Places Density and Adherents Percentage.png*) reveals dense clusters in the Midwest and South, reinforcing the cultural connection between religious framing and supernatural reporting.

Similarly, our **Binge Drinking Correlation** map (*Correlation Between Haunted Places Density and Binge Drinking Rate.png*) displays moderate overlap in areas like central Ohio and Missouri, suggesting possible ties between social behavior and heightened haunting reports.

Through the **Haunted Places by Apparition Type** map, we confirmed that ghost sightings (marked in blue) dominate the records, but orbs, UFOs, and other entities still contribute to regional variation (*Haunted Places by Apparition Type.png*).

Our **Average Daylight Hours Choropleth** (*Haunted Places by Average Daylight Hours.png*) indicates that the southern U.S. experiences more hauntings despite longer daylight, while **Time of Day Map** (*Haunted Places by Time of Day.png*) surprisingly reveals that many hauntings occur in the morning. This suggests either a perceptual bias or possible data misclassification.

A **Heatmap of Haunted Places** (*Heatmap for Haunted Places.png*) further confirms strong clustering in high-density areas like Southern California and the Northeast. Finally, the **Normalized Density × Avg Daylight Hours** overlay (*Normalized Density x Avg Daylight Hours.png*) highlights how hauntings persist where historic population centers and moderate daylight intersect—particularly in Pennsylvania and Michigan.

Better analysis should not be limited to a specific set of methods. While we initially applied the tools and techniques outlined in the assignment documentation, we recognized that this structured and multidimensional dataset had much more potential. As such, we expanded our approach by exploring additional geo-visualization techniques to derive deeper insights:

- When analyzing haunted places by apparition type, we observed that 'Ghost' was by far the most frequently mentioned, which aligns with cultural expectations.
- However, a surprising insight emerged when plotting hauntings by time of day: the majority of sightings were reported in the **morning**, not in the evening or at dusk as one might assume. This

unexpected trend raised the possibility of a data extraction error or inconsistent labeling during our feature engineering phase.
- Our heatmaps showed the most intense haunting clusters in densely populated areas such as Southern California and New England. These regional hot spots likely contribute to stronger correlations when cross-analyzing haunted density with factors like **Adherents (%)** and **Binge Drinking Rate**.

Although these findings were insightful, we still felt our work only scratched the surface. There are undoubtedly more disciplines, creative directions, and analytical techniques that could be applied to this dataset. The opportunities for exploration within data science--especially when working with rich, open-ended datasets like this--are limitless.

Geo-visualization also revealed some underlying data quality issues from our earlier assignments. For example, when mapping haunted place occurrences by time of day, a large majority were marked as "Unknown," suggesting limited or inconsistent feature extraction during Assignment 1. Similarly, when displaying haunted places by average daylight hours, we noticed outliers in states like Alaska, where daylight hour values were unusually high, and clusters in Wyoming where the daylight hour attribute was missing entirely. These anomalies imply potential issues either in how we extracted or joined the daylight data, or inconsistencies in the original external source.

Additionally, some haunted place entries appeared in international locations such as the UK, India, and Japan, indicating a few records with inconsistent or unexpected geotagging beyond U.S. borders.

Regarding the GeoParser experience, we found it far more difficult to implement than anticipated. The application depends on outdated versions of Python and Django, making deployment problematic in modern environments. Server communication was occasionally unclear, and documentation was limited. Interestingly, the built-in GeoParser accuracy--especially through Tika--was noticeably subpar in our tests.

Given that GeoParser uses the same underlying logic as Tika, we believe a more robust alternative would be to independently run location extraction using tools like Lucene-Geo-Gazetteer and Apache Tika, then apply spatial analysis via modern external geo-packages or visualization tools such as QGIS or Leaflet. This approach would likely offer more control, transparency, and extensibility for future work.

4. Thoughts on ImageSpace and ImageCat

What was easy:

- Installation and documentation for ImageSpace was clear.
- The image upload and PostgreSQL ingestion succeeded.

What was difficult:

- A few of the ImageSpace files (docker-compose.yml, import-images.sh) needed to be tweaked in order to (almost) properly set up the interface.
- Thumbnails didn't show up on the interface, making analysis difficult.
- Similarity UI failed with 500 errors and infinite spinner multiple times before tweaking files solved this problem.

Recommendation: Implement some of the GitHub pull requests (especially 199 and 203) to better explain the setup process, as well as some checks to ensure everything is running well throughout.

5. Notes on Scripts and JSON Conversion

We wrote 4 main Python scripts for TSV preprocessing and JSON conversion:

1. **prepare_data.py** / **prepare_data_100.py** – clean and sample top N haunted places by core attributes.
2. **aggregate_by_year.py** – group hauntings by year of nearest historical place.
3. **aggregate_parallel_data.py** – compute summary stats for parallel coordinates plot (hauntings, crime rate, historical age).

Each script outputs a D3-compatible JSON file and handles missing data gracefully using **pandas**, **os**, and **json**.

Appendix: Challenges Summary (Task 4)

- Image thumbnail would not appear on the ImageSpace website
- Docker network conflicts required YML file edits and manual cleanup
- Using the default imagespace-network block meant Solr, Girder, and other services never came up under the expected deploy_imagespace-network. Had to tweak the YML file.

Despite these setbacks, we preserved image assets and captured project completeness through alternative visual means.